# Predicting Election Trends By Web Data

Kaan Cem Ketenci

Deji Suolang

# Research Questions

Can we predict dynamic changes in voter support of political candidates by social media and web data?

- How does predictions informed by real-time web-data map with existing benchmark national opinion polls?
- Is web data informative of shifts in public opinion before, during and after debate?

Focus: October 15, 2019 Democratic Debate

Larger volume of social media data/post/search volume/opinion shifts

# Our study will use four primary data sources:

1. **Twitter streaming API**

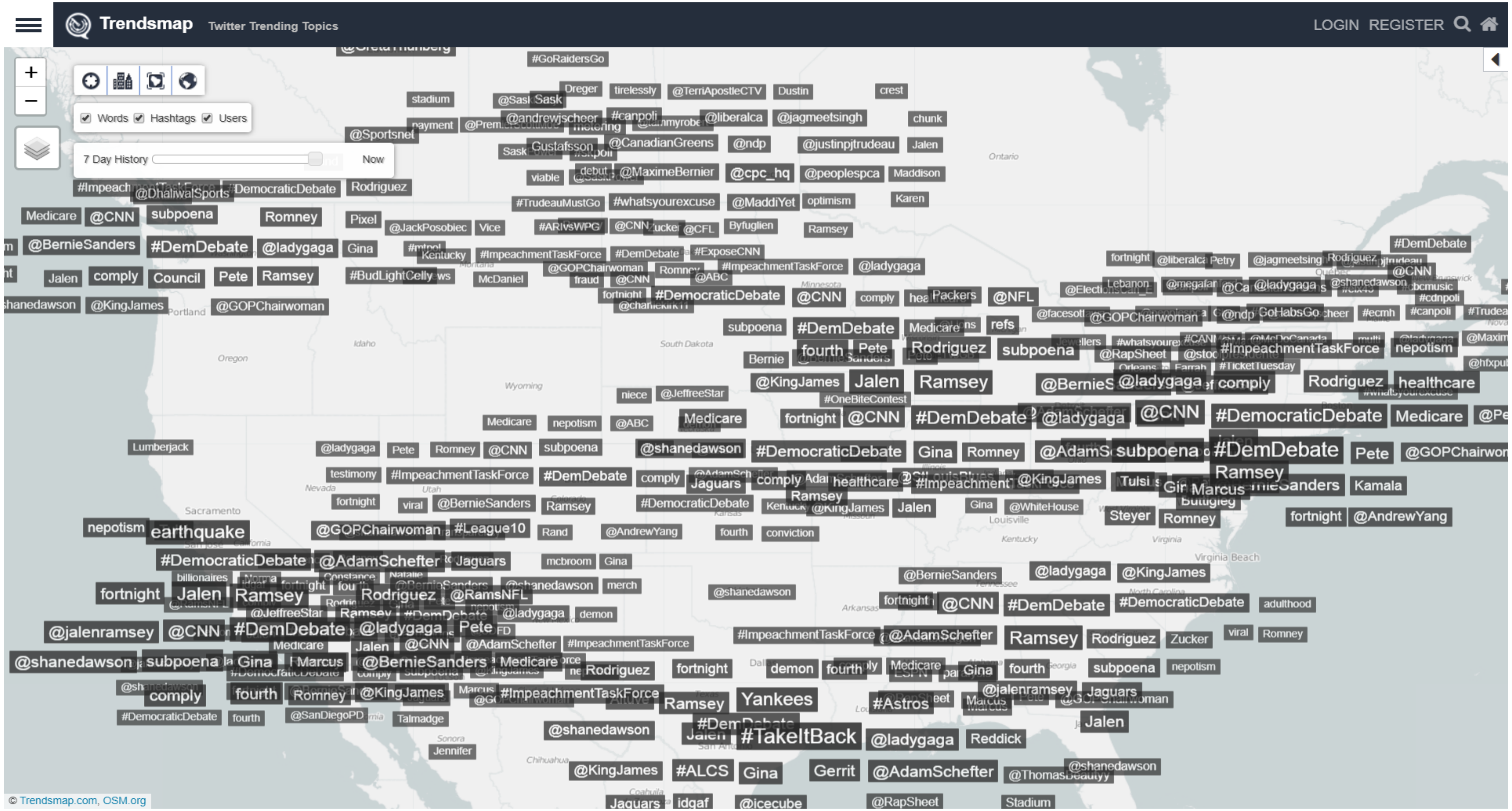
2. **2016 American Community Survey Data**

3. **Google Trends**

4. **Odd checkers Website**

# Twitter Trendsmap of Trending Topics During the Debate

https://www.trendsmap.com/

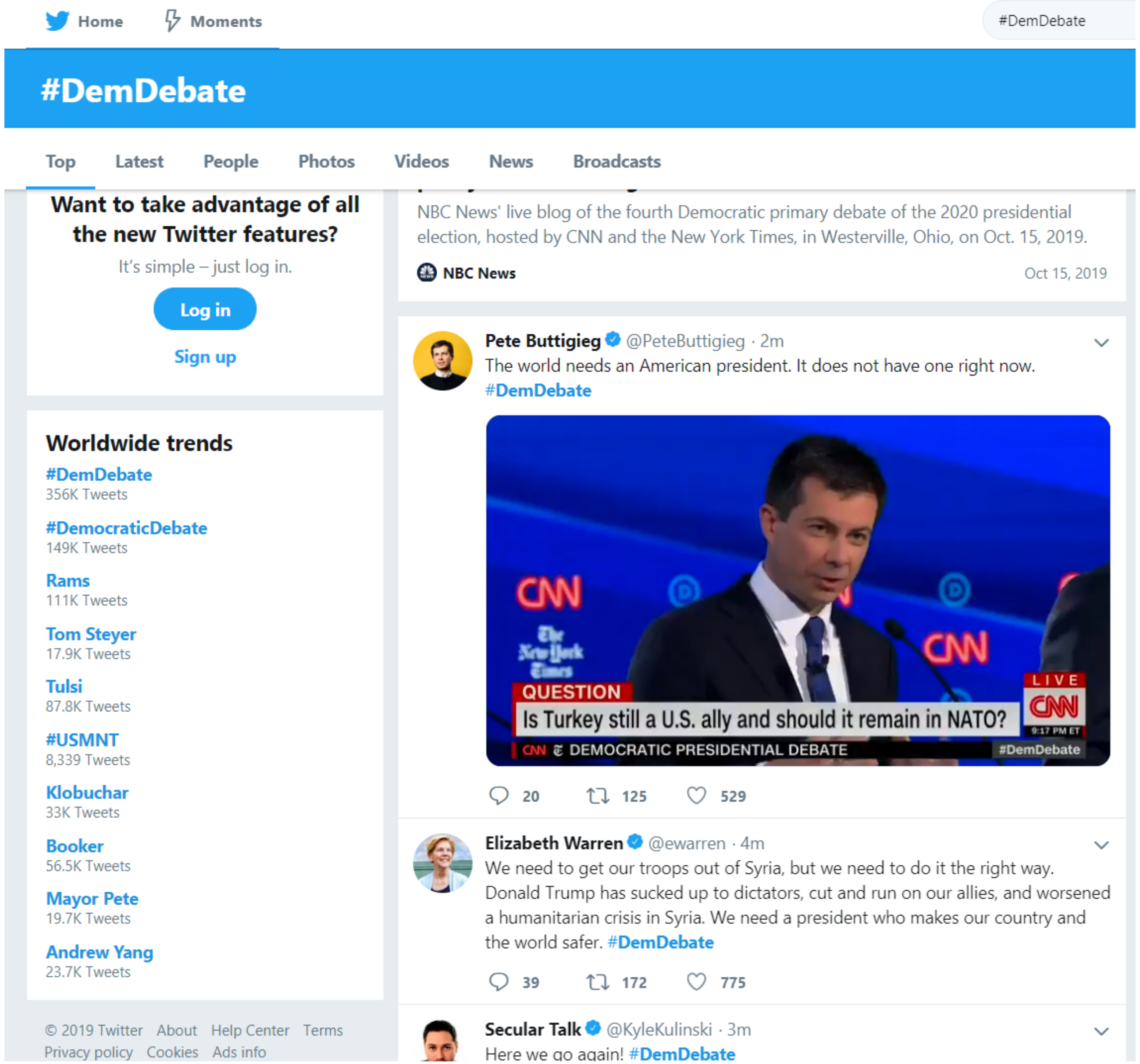


☑ Words ☑ Hashtags ☑ Users

7 Day History Now

# Data source 1: Twitter streaming API

- Listening tweets from October 10 to October 20, 2019 for 10 days.
- Hashtags/Keywords: names and related terms of 12 candidates
- Tracking the number of twitter follower during the debate



Home Moments #DemDebate

#DemDebate

Top Latest People Photos Videos News Broadcasts

Want to take advantage of all the new Twitter features?

It's simple – just log in.

Log in

Sign up

Worldwide trends

#DemDebate
356K Tweets

#DemocraticDebate
149K Tweets

Rams
111K Tweets

Tom Steyer
17.9K Tweets

Tulsi
87.8K Tweets

#USMNT
8,339 Tweets

Klobuchar
33K Tweets

Booker
56.5K Tweets

Mayor Pete
19.7K Tweets

Andrew Yang
23.7K Tweets

© 2019 Twitter About Help Center Terms
Privacy policy Cookies Ads info

NBC News' live blog of the fourth Democratic primary debate of the 2020 presidential election, hosted by CNN and the New York Times, in Westerville, Ohio, on Oct. 15, 2019.

NBC News Oct 15, 2019

Pete Buttigieg @PeteButtigieg · 2m
The world needs an American president. It does not have one right now. #DemDebate

CNN
The New York Times
QUESTION
Is Turkey still a U.S. ally and should it remain in NATO?
LIVE
CNN
9:17 PM ET
CNN DEMOCRATIC PRESIDENTIAL DEBATE
#DemDebate

20 125 529

Elizabeth Warren @ewarren · 4m
We need to get our troops out of Syria, but we need to do it the right way. Donald Trump has sucked up to dictators, cut and run on our allies, and worsened a humanitarian crisis in Syria. We need a president who makes our country and the world safer. #DemDebate

39 172 775

Secular Talk @KyleKulinski · 3m
Here we go again! #DemDebate

# Twitter streaming API (cont’d)

**Sample tweets**

As we obtained a very large json file (100 GB in total), we randomly sampled 10000 for each of 12 candidate per day

**Segregating positive and negative tweets**

- Cleaning: Select relevant variables of the tweets such as “text”, “retweet text” and “location”, remove hashtags and URLs and other twitter handles.
- Use get_sentiment() function to extract sentiment score for each of the tweets to understand the change over time.
- Classify the tweet sentiment as positive, negative and neutral(1, -1 and 0).

**Volumetric Information**

Changes in the number of followers and the public support of candidates may be correlated.

# Data Source 2: 2016 American Community Survey Data

- Use the “censusapi” R package to interface with the Census Bureau’s API to streamline the data collection process

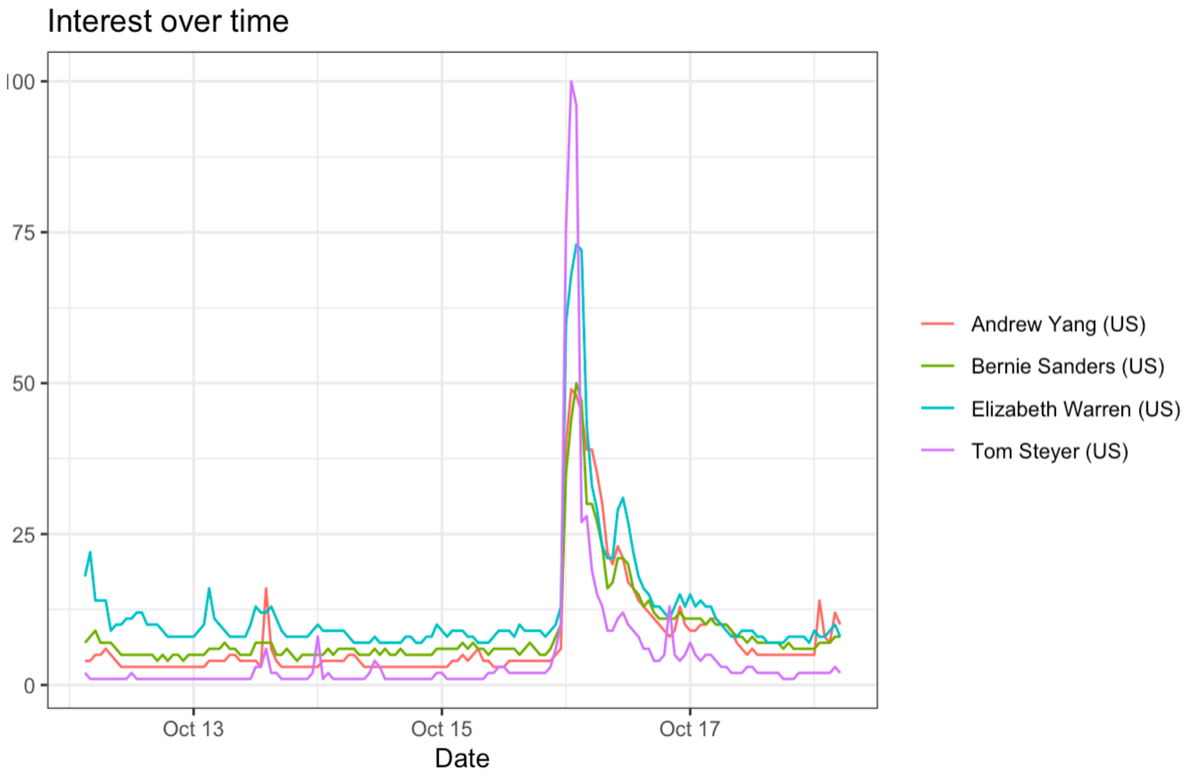
- Variable selections: Select socio-demographic information at state level from 2016 ACS including “age”, “household income” and racial information to compute “Black African American Ratio”.

- Data cleaning and recoding, for example, recode “66666” into NA

- Use “state name” variable to link with twitter data and google trends data for further analysis.

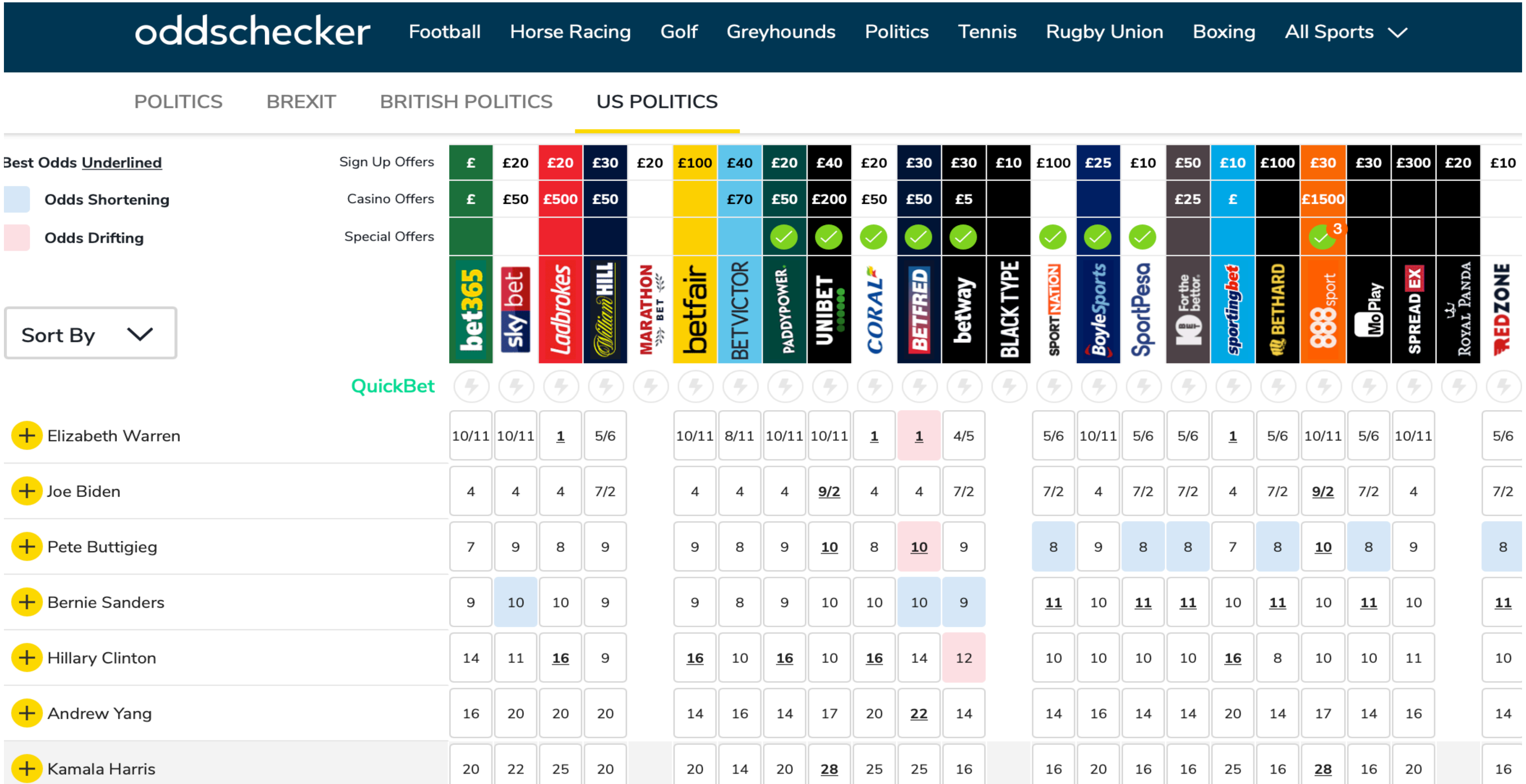
# Data Source 3: Google Trend

- Webscraping of daily and hourly number of Google searches of each of the 12 candidates’ names between October 10 and October 20.
- This Google search statistic may primarily reflect dynamic shifts in popularity of each candidate and how often people are talking about them. It does not involve any sentiment analysis or categorization of searches into positve or negative searches for candidates.



- As expected, there is a big spike in the hourly searches in the evening of October 15th during the debate hours and several hours following the debate.

# Data Source 4: Odds Checker Website



oddschecker Football Horse Racing Golf Greyhounds Politics Tennis Rugby Union Boxing All Sports

POLITICS BREXIT BRITISH POLITICS US POLITICS

Best Odds Underlined

Odds Shortening

Odds Drifting

Sort By

| | bet365 | sky bet | Ladbrokes | William HILL | MARATHON BET | betfair | BETVICTOR | PADDYPOWER. | UNIBET | CORAL | BETFRED | betway | BLACK TYPE | SPORT NATION | BoyleSports | SportPesa | 10BET For the bettor. | sportingbet | BETHARD | 888sport | MoPlay | SPREAD EX | ROYAL PANDA | REDZONE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sign Up Offers | £ | £20 | £20 | £30 | £20 | £100 | £40 | £20 | £40 | £20 | £30 | £30 | £10 | £100 | £25 | £10 | £50 | £10 | £100 | £30 | £30 | £300 | £20 | £10 |
| Casino Offers | £ | £50 | £500 | £50 | | | £70 | £50 | £200 | £50 | £50 | £5 | | | | | £25 | £ | | £1500 | | | | |
| Special Offers | | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | ✓ 3 | | | | |
| QuickBet | | | | | | | | | | | | | | | | | | | | | | | | |
| Elizabeth Warren | 10/11 | 10/11 | 1 | 5/6 | | 10/11 | 8/11 | 10/11 | 10/11 | 1 | 1 | 4/5 | | 5/6 | 10/11 | 5/6 | 5/6 | 1 | 5/6 | 10/11 | 5/6 | 10/11 | | 5/6 |
| Joe Biden | 4 | 4 | 4 | 7/2 | | 4 | 4 | 4 | 9/2 | 4 | 4 | 7/2 | | 7/2 | 4 | 7/2 | 7/2 | 4 | 7/2 | 9/2 | 7/2 | 4 | | 7/2 |
| Pete Buttigieg | 7 | 9 | 8 | 9 | | 9 | 8 | 9 | 10 | 8 | 10 | 9 | | 8 | 9 | 8 | 8 | 7 | 8 | 10 | 8 | 9 | | 8 |
| Bernie Sanders | 9 | 10 | 10 | 9 | | 9 | 8 | 9 | 10 | 10 | 10 | 9 | | 11 | 10 | 11 | 11 | 10 | 11 | 10 | 11 | 10 | | 11 |
| Hillary Clinton | 14 | 11 | 16 | 9 | | 16 | 10 | 16 | 10 | 16 | 14 | 12 | | 10 | 10 | 10 | 10 | 16 | 8 | 10 | 10 | 11 | | 10 |
| Andrew Yang | 16 | 20 | 20 | 20 | | 14 | 16 | 14 | 17 | 20 | 22 | 14 | | 14 | 16 | 14 | 14 | 20 | 14 | 17 | 14 | 16 | | 14 |
| Kamala Harris | 20 | 22 | 25 | 20 | | 20 | 14 | 20 | 28 | 25 | 25 | 16 | | 16 | 20 | 16 | 16 | 25 | 16 | 28 | 16 | 20 | | 16 |

https://www.oddschecker.com/politics/us-politics/us-presidential-election-2020/democrat-candidate

# Web Scraping from Odd checkers Website

- Oddschecker is a website that combines betting odds for various bets from all major betting companies. We are interested in the bets on who will become the Democratic Party US presidential nominee. We focus on 12 candidates who participated in the CNN debate and collected data between October 10 and October 20.
- Convert betting odds data into implied probabilities.
- We collected data once a day. We further collected data shortly before and after the debate.
- The increases and decreases in implied probabilities of winning the nomination can be compared with political analyses by mainstream news organizations of debate performances of each candidate.

# Questions?